Learning joint intensity-depth sparse representations

Ivana Tošić and Sarah Drewes

Abstract

This paper presents a method for learning overcomplete dictionaries composed of two modalities that describe a 3D scene: image intensity and scene depth. We propose a novel Joint Basis Pursuit (JBP) algorithm that finds related sparse features in two modalities using conic programming and integrate it into a two-step dictionary learning algorithm. JBP differs from related convex algorithms because it finds joint sparsity models with different atoms and different coefficient values for intensity and depth. This is crucial for recovering generative models where the same sparse underlying causes (3D features) give rise to different signals (intensity and depth). We give a theoretical bound for the sparse coefficient recovery error obtained by JBP, and show experimentally that JBP is far superior to the state of the art Group Lasso algorithm. When applied to the Middlebury depth-intensity database, our learning algorithm converges to a set of related features, such as pairs of depth and intensity edges or image textures and depth slants. Finally, we show that the learned dictionary and JBP achieve the state of the art depth inpainting performance on time-of-flight 3D data.

Index Terms

Sparse approximations, dictionary learning, hybrid image-depth sensors.

I. Introduction

Hybrid image-depth sensors have recently gained a lot of popularity in many vision applications. Time of flight cameras [1], [2] provide real-time depth maps at moderate spatial resolutions, aligned with the image data of the same scene. Microsoft Kinect [3] also provides real-time depth maps that can be registered with color data in order to provide 3D scene representation. Since captured images and depth data are caused by the presence of same objects in a 3D scene, they represent two modalities of the same phenomena and are thus correlated. This correlation can be advantageously used for denoising corrupted or inpainting missing information in captured depth maps. Such algorithms are of significant importance to technologies relying on image-depth sensors for 3D scene reconstruction or visualization [3], [4], where depth maps are usually noisy, unreliable or of poor spatial resolution.

Solving inverse problems such as denoising or inpainting usually involves using prior information about data. Sparse priors over coefficients in learned linear generative models have been recently applied to these problems with large success [5]–[7]. A similar approach has been proposed for learning sparse models of depth only, showing state-of-the-art performance in depth map denoising and offering a general tool for improving existing depth estimation algorithms [8]. However, learning sparse generative models for joint representation of depth and intensity images has not been addressed yet. Learning such models from natural 3D data is of great importance for many applications involving 3D scene reconstruction, representation and compression.

This paper proposes a method for learning joint depth and intensity generative models. Each of these two modalities is represented using overcomplete linear models, resulting in two sets of coefficients. These two sets are coupled via a set of hidden variables, where each variable multiplies exactly one coefficient in each modality. Consequently, imposing a sparse prior on this set of coupling variables results in a common sparse support of intensity and depth. Each of these hidden variables can be interpreted as presence of a depth-intensity feature pair arising from the same underlying cause in a 3D scene. To infer these hidden variables under a sparse prior, we propose a convex, second order cone program named Joint Basis Pursuit (JBP). Compared to Group Lasso (GL) [9], which is commonly used for coupling sparse variables, JBP gives significantly smaller coefficient inference error. In addition, we derive theoretical bounds for the coefficient recovery error in JBP, exploiting the restricted isometry

I. Tošić is with Ricoh Innovations, Inc., Menlo Park, USA, email: ivana@rii.ricoh.com. This work has been performed while she was with the Helen Wills Neuroscience Institute, University of California, Berkeley, USA. S. Drewes is with T-Systems International GmbH, Darmstadt, Germany, sarah.drewes@t-systems.com. She performed this work while she was with the Department of Industrial Engineering and Operations Research at University of California, Berkeley.

This work has been supported by the Swiss National Science Foundation under the fellowship PA00P2-134159 awarded to I. Tošić.

property (RIP) [10] of the model. Finally, we propose an intensity-depth dictionary learning algorithm based on the new model and JBP. We show its superiority to GL in model recovery experiments using synthetic data, as well as in inpainting experiments using real time-of-flight 3D data.

Section II introduces the proposed intensity-depth generative model. Inference of its hidden variables is achieved via the new JBP algorithm presented in Section III, while learning of model parameters is explained in Section IV. Section V gives relations of the proposed JBP to prior art. Experimental results are presented in Section VI.

II. Intensity-depth generative model

Before introducing the proposed intensity-depth generative model, let us set the notation rules. Throughout the rest of the paper, vectors are denoted with bold lower case letters and matrices with bold upper case letters. Letters I, D in superscripts refer to intensity and depth, respectively. Sets are represented with calligraphic fonts. Column-wise and row-wise concatenations of vectors $\mathbf{a}$ and $\mathbf{b}$ are denoted as $[\mathbf{a}\ \ \mathbf{b}]$ and $[\mathbf{a}; \mathbf{b}]$, respectively.

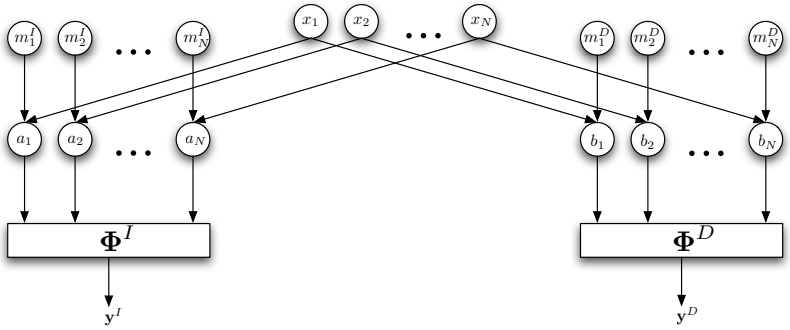


Fig. 1. Graphical representation of the proposed intensity-depth generative model.

Let us now explain the proposed joint depth-intensity generative model whose graphical representation is shown in Figure 1. Intensity image $\mathbf{y}^I$ and depth image $\mathbf{y}^D$ (in vectorized forms) are assumed to be sparse in dictionaries $\mathbf{\Phi}^I$, resp. $\mathbf{\Phi}^D$, i.e. they are represented as linear combinations of dictionary elements:

$$\begin{aligned}\mathbf{y}^I &= \mathbf{\Phi}^I\mathbf{a} + \boldsymbol{\eta}^I = \sum_{i\in\mathcal{I}_0} \boldsymbol{\phi}_i^I a_i + \boldsymbol{\eta}^I \\ \mathbf{y}^D &= \mathbf{\Phi}^D\mathbf{b} + \boldsymbol{\eta}^D = \sum_{i\in\mathcal{I}_0} \boldsymbol{\phi}_i^D b_i + \boldsymbol{\eta}^D,\end{aligned} \tag{1}$$

where vectors $\mathbf{a}$ and $\mathbf{b}$ have a small number of non-zero elements and $\boldsymbol{\eta}^I$ and $\boldsymbol{\eta}^D$ represent noise vectors. $\mathcal{I}_0$ is the set of indexes identifying the columns (i.e., atoms) of $\mathbf{\Phi}^I$ and $\mathbf{\Phi}^D$ that participate in sparse representations of $\mathbf{y}^I$ and $\mathbf{y}^D$. Its cardinality is much smaller than the dictionary size, hence $|\mathcal{I}_0| \ll |\mathcal{I}|$, where $\mathcal{I} = \{1, 2, ..., N\}$ denotes the index set of all atoms. This means that each image can be represented as a combination of few, representative image features described by dictionary elements, modulated by their respective coefficients. Because depth and intensity features correspond to two modalities arising from the same 3D features, we model the coupling between coefficients a_i and b_i through latent variables x_i as:

$$\begin{aligned}a_i &= m_i^I x_i \\ b_i &= m_i^D x_i, \quad \forall i \in \mathcal{I},\end{aligned} \tag{2}$$

where the variables m_i^I, m_i^D represent the magnitudes of the sparse coefficients and x_i represent the activity of these coefficients. Ideally, these variables should be binary where 0 represents the absence and 1 represents the presence of a pair of depth-intensity features. In that case $\sum_i x_i$ counts the number of non-zero pairs of depth-intensity features. However, inference of binary values represents a combinatorial optimization problem of high complexity which depends on dictionary properties and the permission of noise, cf. [11]. We relax the problem by allowing x_i to attain continuous values between 0 and 1, which has been proven to provide a very good approximation in a similar context, cf., e.g., [12], [13].

By assuming that the vector $\mathbf{x} = (x_1, x_2, ..., x_N)^\mathsf{T}$ is sparse, we assume that $\mathbf{y}^I$ and $\mathbf{y}^D$ are described by a small number of feature pairs $(\boldsymbol{\phi}_i^I, \boldsymbol{\phi}_i^D)$ that are either prominent in both modalities (both m_i^I and m_i^D are significant) or in only one modality (either m_i^I or m_i^D is significant). In these cases x_i is non-zero, which leads to non-zero values for either a_i or b_i, or both. If x_i is zero or very close to zero, both a_i and b_i are also zero or very small. Hence, the sparsity assumption on $\mathbf{x}$ enforces a compact description of both modalities. It uses simultaneously active coefficients in both modalities, or when such pairs cannot approximate both images, the model also allows only one coefficient within a pair to be non-zero. Therefore, the model represents intensity and depth using a small set of joint features and a small set of independent features. The main challenge is to simultaneously infer the latent variables $\mathbf{x}$, $\mathbf{m}^I = (m_1^I, m_2^I, ..., m_N^I)^\mathsf{T}$ and $\mathbf{m}^D = (m_1^D, m_2^D, ..., m_N^D)^\mathsf{T}$ under the sparsity assumption on $\mathbf{x}$. In the next section we propose to solve this problem via convex optimization.

III. Joint Basis Pursuit

Let us re-write the intensity-depth generative model, including all latent variables, in matrix notation as:

$$\begin{bmatrix} \mathbf{y}^I \\ \mathbf{y}^D \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Phi}^I & 0 \\ 0 & \boldsymbol{\Phi}^D \end{bmatrix} \cdot \begin{bmatrix} \mathbf{M}^I \\ \mathbf{M}^D \end{bmatrix} \cdot \mathbf{x} + \begin{bmatrix} \boldsymbol{\eta}^I \\ \boldsymbol{\eta}^D \end{bmatrix},$$

where $\mathbf{M}^I = \text{diag}(m_1^I, m_2^I, ..., m_N^I)$ and $\mathbf{M}^D = \text{diag}(m_1^D, m_2^D, ..., m_N^D)$. Suppose that we know the dictionaries $\boldsymbol{\Phi}^I$ and $\boldsymbol{\Phi}^D$ and we want to find joint sparse representations of intensity and depth, i.e., to infer latent variables $\mathbf{x}, \mathbf{m}^I, \mathbf{m}^D$. To do this, we need to solve the following optimization problem:

$$\text{OPT1}: \min \sum_i x_i, \;\; \text{where } x_i \in [0,1],\; i = 1, ..., N$$

$$\begin{aligned} \text{subject to: } & \|\mathbf{y}^I - \boldsymbol{\Phi}^I \mathbf{M}^I \mathbf{x}\|^2 \leq (\epsilon^I)^2 && (3)\\ & \|\mathbf{y}^D - \boldsymbol{\Phi}^D \mathbf{M}^D \mathbf{x}\|^2 \leq (\epsilon^D)^2 && (4)\\ & |m_i^I| \leq U^I && (5)\\ & |m_i^D| \leq U^D && (6) \end{aligned}$$

where ϵ^I, ϵ^D are allowed approximation errors and U^I and U^D are upper bounds on the magnitudes $\mathbf{m}^I$ and $\mathbf{m}^D$. In practice, the values of these upper bounds can be chosen as arbitrarily high finite values. This optimization problem is hard to solve using the above formulation, since the first two constraints are non-convex due to the terms $\mathbf{M}^I\mathbf{x}$ and $\mathbf{M}^D\mathbf{x}$ which are bilinear in the variables $\mathbf{x}$, $\mathbf{m}^I$ and $\mathbf{m}^D$. To overcome this issue, we transform it into an equivalent problem by introducing the change of variables given by Eqs. (2) deriving:

$$\text{OPT2}: \min \sum_i x_i, \;\; \text{where } x_i \in [0,1],\; i = 1, ..., N$$

$$\begin{aligned} \text{subject to: } & \|\mathbf{y}^I - \boldsymbol{\Phi}^I \mathbf{a}\|^2 \leq (\epsilon^I)^2 && (7)\\ & \|\mathbf{y}^D - \boldsymbol{\Phi}^D \mathbf{b}\|^2 \leq (\epsilon^D)^2 && (8)\\ & |a_i| \leq U^I x_i && (9)\\ & |b_i| \leq U^D x_i, && (10) \end{aligned}$$

which is a convex optimization problem with linear and quadratic constraints that can be solved efficiently, i.e., in polynomial time, using log-barrier algorithms, cf. [14], [15]. A variety of free and commercial software packages are available like IBM ILOG CPLEX [16], that we use in our experiments.

The problems (OPT1) and (OPT2) are indeed equivalent using the variable transformation in Eqs. (2) as follows.

Lemma 1. *For any optimal solution* $(\mathbf{x}^*, \mathbf{a}^*, \mathbf{b}^*)$ *of (OPT2),* $\mathbf{x}^*$ *is also an optimal solution to (OPT1) with corresponding matrices* $(\mathbf{M^I})^*$, $(\mathbf{M^D})^*$ *according to* (2).
Also, any optimal solution $(\mathbf{x}^*, (\mathbf{M^I})^*, (\mathbf{M^D})^*)$ *of (OPT1) defines an optimal solution* $(\mathbf{x}^*, \mathbf{a}^*, \mathbf{b}^*)$ *to (OPT2) .*

Proof: For any $(\mathbf{x}^*, \mathbf{a}^*, \mathbf{b}^*)$ and corresponding $(\mathbf{M^I})^*$, $(\mathbf{M^D})^*$ that satisfy Eqs. (2), conditions (7) and (8) are equivalent to (3) and (4) by definition. Moreover, since $\mathbf{x}^*$ is nonnegative, conditions (9) and (10) are equivalent

to (5) and (6). Hence, any $\mathbf{x}^*$ that is optimal for (OPT2) with corresponding $(\mathbf{a}^*, \mathbf{b}^*)$ is optimal for (OPT1) with corresponding $(\mathbf{M^I})^*$, $(\mathbf{M^D})^*$ and vice versa. ■

An immediate consequence of the form of the objective function and constraints in (OPT2) is that $\mathbf{x}^*$ is chosen such that (9) and (10) are both feasible and at least one of them is active. Formally, this is stated by the following lemma.

Lemma 2. *For any optimal solution $(\mathbf{x}^*, \mathbf{a}^*, \mathbf{b}^*)$ of (OPT2), at least one of the constraints* (9) *and* (10) *is active for each component i, hence we have*

$$x_i^* = \max\{\frac{|a_i^*|}{U^I}, \frac{|b_i^*|}{U^D}\}, \quad \forall i = 1, ..., N. \tag{11}$$

Proof: Otherwise it would be a contradiction to the optimality of $\mathbf{x}^*$. ■

In the following, we refer to the optimization problem (OPT2) as Joint Basis Pursuit (JBP), where $\mathbf{x}$ is the vector of joint (coupling) variables in the signal model. It is important to know the theoretical bounds on the norm of the difference between the solution $(\mathbf{a}^*, \mathbf{b}^*)$ found by JBP and the true coefficients $(\mathbf{a}, \mathbf{b})$ of the model (1).

Based on the non-coupled case that is treated in [11], we develop bounds on the difference of the optimal solution of (OPT2) and a sparse signal to be recovered. For this purpose, we assume that the matrix

$$\mathbf{A} := \begin{bmatrix} \mathbf{\Phi^I} & \mathbf{0} \\ \mathbf{0} & \mathbf{\Phi^D} \end{bmatrix} \tag{12}$$

satisfies the restricted isometry property with a constant δ_S, which refers to the following characteristics of the linear system. Denote $\mathbf{A}_T$, $T \subset 1, ..., n$ as the $n \times |T|$ submatrix obtained by extracting the columns of $\mathbf{A}$ corresponding to the indices in T. The S-restricted isometry constant δ_S is then defined as:

Definition 1. *[10] The S-restricted isometry constant δ_S of $\mathbf{A}$ is the smallest quantity such that*

$$(1 - \delta_S)\|\mathbf{s}\|_2^2 \leq \|\mathbf{A}_T\mathbf{s}\|_2^2 \leq (1 + \delta_S)\|\mathbf{s}\|_2^2 \tag{13}$$

for all subsets T with $|T| \leq S$ and coefficient sequences (s_j), $j \in T$.

When $\delta_S << 1$, this property requires that every set of columns with cardinality less than S approximately behaves like an orthonormal system. It can thus be related to the maximal value of the inner product between any two columns in the matrix $\mathbf{A}$, usually called the coherence of the dictionary:

$$\mu = \max_{i,j \neq i} |\langle \phi_i, \phi_j \rangle|, \tag{14}$$

where ϕ_i and ϕ_j are two different atoms in the dictionary (i.e., two columns of $\mathbf{A}$) and $\langle \cdot \rangle$ denotes the inner product. With this definition, it can be easily shown that $\delta_S = \mu(|T| - 1)$ satisfies the RIP inequality (13).

Before we present the bound on the coefficient recovery error of JBP, let us first define some prerequisites. Assume we are given a pair of sparse signals $(\mathbf{y}^I, \mathbf{y}^D)$ as in Eq. (1), with sparse coefficients $(\mathbf{a}^0, \mathbf{b}^0)$, which satisfy constraints (7) and (8). Let T_0 be the support of $\mathbf{x}^0$ which is at the same time the support of at least $\mathbf{a}^0$ or $\mathbf{b}^0$ and contains the support of the other one or it coincides with the support of both. Without loss of generality, let us assume that

$$\|\mathbf{y}^I\|_2 = \|\mathbf{y}^D\|_2 =: f_0, \tag{15}$$

which can be easily obtained by normalization. Assume also that the components of $\mathbf{a}^0$ and $\mathbf{b}^0$ satisfy the bound constraints[1]

$$|a_i^0| \leq f_0, \;\; |b_i^0| \leq f_0, \; \forall i = 1, ..., N, \tag{16}$$

i.e., in the remainder of the paper we assume the same bounds on a_i and b_i: $U^I = U^D = U = f_0$. It is also useful in practice to select the approximation error ϵ in terms of the fraction of the total signal energy, so we denote $\epsilon = \eta f_0$, where $0 \leq \eta < 1$.

[1]Although the assumption in Eq. (16) does not hold in general, in practical applications using learned dictionaries we found that it is always satisfied. However, if one wants to use a bound that is always satisfied, one should choose $U = f_0/\sigma_{min}$, where σ_{min} is the smallest of all singular values of $\mathbf{\Phi}^I$ and $\mathbf{\Phi}^D$.

Let further α_i denote the scale between the smaller and larger coefficient for each index i within the sparse support set T_0, i.e.:

$$\alpha_i = \min\{\frac{|a_i^0|}{|b_i^0|}, \frac{|b_i^0|}{|a_i^0|}\}, \quad \forall i \in T_0, \tag{17}$$

and let γ denote:

$$\gamma = 1 - \min_{i \in T_0} \alpha_i. \tag{18}$$

Parameter γ describes the level of similarity between the sparse coefficients in the two signals, which is decreasing with higher similarity. In the trivial case when $a_i^0 = b_i^0$ for all $i \in T_0$ we have that $\gamma = 0$. In all other cases $\gamma \leq 1$. Due to the common sparse support of intensity and depth as assumed by the signal model, we usually have $a_i^0 \neq 0$ and $b_i^0 \neq 0$, for all $i \in T_0$, and hence $\gamma < 1$.

Let further x^0 denote an auxiliary vector that satisfies

$$\max\{|a_i^0|, |b_i^0|\} = U x_i^0,$$

namely (x^0, a^0, b^0) is a feasible solution to (OPT2), where x^0 is chosen such that (9) and (10) are both feasible and (at least) one of them is active.

Finally, let (x^*, a^*, b^*) be an optimal solution to (OPT2). Then we have the following worst case bound on the distance of these.

Theorem 1. *Let (a^0, b^0) and (a^*, b^*) as defined above and choose $U = f_0$ with f_0 from* (15) *and $\epsilon^I = \epsilon^D = \eta f_0$, where $0 \leq \eta < 1$. Then*

$$\|[a^0; b^0] - [a^*; b^*]\|_2^2 \leq \left[\frac{|T_0|^2}{M}(C + \gamma)^2 + C^2\right] f_0^2 \tag{19}$$

holds for a constant C that depends on the signal model parameter γ, the sparse support size $|T_0|$, approximation parameter η and where the M-restricted isometry property is satisfied for the linear system, cf. Def. 1. In particular, we have:

$$C = \frac{4\eta\sqrt{M} + \gamma|T_0|\sqrt{1+\delta_M}}{\sqrt{M(1-\delta_{M+|T_0|})} - \sqrt{|T_0|(1+\delta_M)}}. \tag{20}$$

The proof of this Theorem is given in Appendix A.

IV. Intensity-depth dictionary learning

In the previous section we have shown how to find sparse coefficients in the joint depth-intensity generative model, assuming that the model parameters, i.e., dictionaries $\mathbf{\Phi}^I$ and $\mathbf{\Phi}^D$ are given. Since we do not have those parameters in general, we propose to learn them from a large database of intensity-depth image examples. Dictionary learning for sparse approximation has been a topic of intensive research in the last couple of years. Almost all existing algorithms are based on Expectation-Maximization, i.e., they are iterative algorithms that consist of two steps: 1) inference of sparse coefficients for a large set of signal examples while keeping the dictionary parameters fixed, and 2) dictionary optimization to minimize the reconstruction error while keeping the coefficients fixed. We follow the same approach here, using JBP in the first step and then conjugate gradient in the second step. Once JBP finds the sparse coefficients $(\mathbf{a}^*, \mathbf{b}^*)$ and the coupling variables $\mathbf{x}$, optimization of $\mathbf{\Phi}^I$ and $\mathbf{\Phi}^D$ becomes decoupled. Therefore, in the learning step we independently optimize the following objectives:

$$(\mathbf{\Phi}^I)^* = \min_{\mathbf{\Phi}^I} \|\mathbf{Y}^I - \mathbf{\Phi}^I\mathbf{A}\|_F^2 + \rho\|\mathbf{\Phi}^I\|_F \tag{21}$$

$$(\mathbf{\Phi}^D)^* = \min_{\mathbf{\Phi}^D} \|\mathbf{Y}^D - \mathbf{\Phi}^D\mathbf{B}\|_F^2 + \rho\|\mathbf{\Phi}^D\|_F, \tag{22}$$

where $\|\cdot\|_F$ denotes the Frobenius norm, $\mathbf{Y}^I$, $\mathbf{Y}^D$, $\mathbf{A}$ and $\mathbf{B}$ are matrices whose columns are $\mathbf{y}_j^I$, $\mathbf{y}_j^D$, $\mathbf{a}_j$ and $\mathbf{b}_j$ respectively, and $j = 1, ..., J$ indexes the signal examples from a given database. In addition to the reconstruction error, we have added a normalization constraint on the dictionaries, scaled by a small parameter ρ, in order to control the dictionary norms as usually done in dictionary learning. Before showing the performance of the proposed learning algorithm, we review prior art that we will use for experimental comparisons in Section VI.

V. Relation to prior art

To the best of our knowledge, there has not been any work that addresses the problem of learning joint intensity-depth sparse representations. Therefore, we overview prior work that focuses on sparse approximation algorithms that bear similarities to the proposed JBP algorithm. Since the main characteristic of JBP is to find sparse approximations of two signals sharing a common sparse support, we overview algorithms targeting this problem. Such algorithms can be grouped into two categories with respect to the signal model they address: a) simultaneous sparse approximation algorithms, and b) group sparse approximation algorithms. We further discuss how algorithms from each group relate to JBP.

Simultaneous sparse approximation algorithms recover a set of jointly sparse signals modeled as:

$$\mathbf{y}_s = \mathbf{\Phi}\mathbf{x}_s + \boldsymbol{\epsilon}_s = \sum_{i \in \mathcal{I}} \boldsymbol{\phi}_i x_s(i) + \boldsymbol{\epsilon}_s, \quad s = 1, ..., S, \tag{23}$$

where S is the total number of signals $\mathbf{y}_s$, $\mathbf{\Phi}$ is the dictionary matrix and $\boldsymbol{\epsilon}_s$ is a noise vector for signal $\mathbf{y}_s$. Vectors of sparse coefficients $\mathbf{x}_s$ share the same sparsity support set $\mathcal{I}$, i.e., they have non-zero entries at the same positions. One of the earliest algorithms in this group is the Simultaneous Variable Selection (SVS) algorithm introduced by Turlach et. al. [17]. SVS selects a common subset of atoms for a set of signals by minimizing the representation error while constraining the ℓ_1-norm of the maximum absolute values of coefficients across signals. Formally, SVS solves the following problem:

$$\text{(SVS)}: \quad \min \frac{1}{2} \sum_{s=1}^{S} \|\mathbf{y}_s - \mathbf{\Phi}\mathbf{x}_s\|^2, \tag{24}$$

$$\text{subject to:} \quad \sum_i \max\{|x_1(i)|, ..., |x_S(i)|\} \leq \tau, \tag{25}$$

where τ is given. Let $\mathbf{X}$ denote the matrix with $\mathbf{x}_s$, $s = 1, ..., S$ as columns. We can see that the left hand side of the constraint in SVS is obtained by applying the ℓ_∞-norm to rows (to find the largest coefficients for all explanatory variables), followed by applying the ℓ_1-norm to the obtained vector in order to promote sparsity of the support. We denote this norm as $\|\mathbf{X}\|_{\infty,1}$. Versions of the same problem for the unconstrained case and the error-constrained case have been studied by Tropp [18].

To see the relation of SVS to JBP, we use Lemma 2, which allows us to formulate the JBP for the special case of $U^I = U^D$ as:

$$\min: \quad t \tag{26}$$

$$\text{subject to:} \quad \|\mathbf{y}^D - \mathbf{\Phi}^D\mathbf{a}\|^2 \leq \epsilon^2 \tag{27}$$

$$\|\mathbf{y}^I - \mathbf{\Phi}^I\mathbf{b}\|^2 \leq \epsilon^2 \tag{28}$$

$$\sum_i \max\{|a_i|, |b_i|\} \leq t. \tag{29}$$

Therefore, JBP operates on the same $\ell_{\infty,1}$-norm of the coefficient matrix as SVS. However, in contrast to SVS, JBP minimizes the number of non-zero elements in both $\mathbf{a}$ and $\mathbf{b}$ by minimizing $\|[\mathbf{a}\ \ \mathbf{b}]\|_{\infty,1}$ and constraining the approximation error induced by the coefficients. A much more important difference of our work and [17] is that we allow for different sets of atoms for intensity and depth. Thus, in JBP, each signal can be represented using a different dictionary, but with coefficient vectors that share the same positions of non-zero entries. This makes JBP applicable to intensity-depth learning, in contrast to SVS. Finally, we remark here that choosing the objective function as we did allows for a smooth convex representation of the last constraint (29).

Group sparse approximation algorithms recover a signal modeled as:

$$\mathbf{y} = \sum_i \mathbf{H}_i \mathbf{x}_i + \boldsymbol{\epsilon}, \tag{30}$$

where $\mathbf{H}_i$ is a submatrix of a big dictionary matrix $\mathbf{H}$. This model is useful for signals whose sparse support has a group structure, namely when groups of coefficients are either all non-zero or all zero. The first algorithm proposed

for group sparse approximation was a generalization of Lasso, developed by Bakin [9], and later studied by other authors (e.g. Yuan and Lin [19]). Group Lasso refers to the following optimization problem:

$$\text{(GL)} : \min \ \|\mathbf{y} - \sum_i \mathbf{H}_i \mathbf{x}_i\|^2 + \lambda \sum_i \|\mathbf{x}_i\|_p, \tag{31}$$

where $\|\cdot\|_p$ denotes the ℓ_p-norm. The most studied variant of group lasso is for $p = 2$, because it leads to a convex optimization problem with efficient implementations. The group sparsity model can be used to represent intensity-depth signals by considering pairs $(a_i, b_i), i = 1, ..., N$ as groups. In this case, group lasso with $p = 2$ becomes:

$$\text{(GL-ID)} : \min \ \|\mathbf{y}^I - \sum_i \boldsymbol{\phi}_i^I a_i\|^2 + \|\mathbf{y}^D - \sum_i \boldsymbol{\phi}_i^D b_i\|^2 + \lambda \sum_i \sqrt{a_i^2 + b_i^2}$$

The drawback of GL with $p = 2$ is that it penalizes independent components (pairs with only one large coefficient) over the more balanced ones (pairs with similar coefficients). Choosing $p = \infty$ avoids this problem and allows selection of pairs with unbalanced coefficients. In that case the regularizer penalizes the norm $\|[\mathbf{a}\ \mathbf{b}]\|_{\infty,1}$. Rather than solving the unconstrained problem of group lasso with $p = \infty$ and a non-smooth objective, JBP reaches a similar goal by solving a constrained convex optimization problem with smooth constraints. It also eliminates the need for tuning the Lagrange multiplier.

VI. Experimental results

We have performed two sets of experiments in order to evaluate the proposed JBP and dictionary learning based on JBP. The first set of experiments uses simulated random data, with the goal to determine the model recovery performance of JBP when the ground truth signal models are given. In the second set, we apply JBP and dictionary learning on real depth-intensity data and show its performance on a depth inpainting task. In both cases, JBP has been compared to Group Lasso (GL).

A. Model recovery

To evaluate the performance of JBP, we have generated multiple sets of pairs of signals of size $N = 64$, denoted by $\{\mathbf{y}_j^I\}$ and $\{\mathbf{y}_j^D\}$, $j = 1, 500$. Signals in each pair have a common sparsity support of size $|T_0|$, and they are sparse in random, Gaussian iid dictionaries $\boldsymbol{\Phi}^I$ and $\boldsymbol{\Phi}^D$ of size 64×128. Their coefficients, $\{\mathbf{a}_j\}$ and $\{\mathbf{b}_j\}$, $j = 1, 500$ are random, uniformly distributed, and do not have the same values nor signs. However, their ratios α_i (as defined in Eq. 17) are bounded from below, which gives a certain value of γ (see Eq. 18) for each set. Hence, we assume some similarity in the magnitudes within each pair of coefficients of the two modalities. All signals have been corrupted by Gaussian noise.

Figure 2 shows the relative coefficient reconstruction error $\|\mathbf{a}^* - \mathbf{a}\|_2^2/\|\mathbf{a}\|_2^2 + \|\mathbf{b}^* - \mathbf{b}\|_2^2/\|\mathbf{b}\|_2^2$, where $(\mathbf{a}^*, \mathbf{b}^*)$ are the reconstructions of original values $(\mathbf{a}, \mathbf{b})$. The error is averaged over 50 different signals and plotted versus the signal-to-noise (SNR) ratio between sparse signals and Gaussian noise. Each of the four graphs is obtained for different values of parameters $|T_0|$ and γ.

We have compared JBP with GL and with the theoretical bounds derived in Section III, for $M = 25$ and $M = 64$. Instead of using the dictionary coherence value for δ, which would give the worst-case bounds, we use the mean of inner products between all atoms to obtain and plot the average case bounds. We can see that JBP always outperforms GL for a large margin. Moreover, the actual performance of JBP is much better than predicted by the theory, showing that the average derived bound is rather conservative.

Furthermore, we have used these randomly generated signals as training sets in our dictionary learning algorithm, in order to recover the original dictionary. For four different values of sparsity $|T_0| = 2, 4, 6, 8$, we have applied the proposed learning algorithm starting from a random initial dictionary. For comparison, we have replaced the JBP in the inference step with GL, while keeping the learning step exactly the same. We refer to this method as GL-based learning. Figure 3(a) shows the mean square error (MSE) between the original atoms and the recovered ones vs sparsity $|T_0|$, for JBP and GL-based learning and two values of γ: 0.1 and 0.25. Similarly, we plot in Figure 3(b) the percentage of recovered atoms vs sparsity, where an atom is considered recovered when its MSE is less than 0.05. Below this threshold the comparison is impossible since GL recovery error is huge (almost 0 recovered atoms). We can see from both graphs that learning based on JBP is far superior to GL and that JBP is less sensitive to values of γ than GL-based learning.

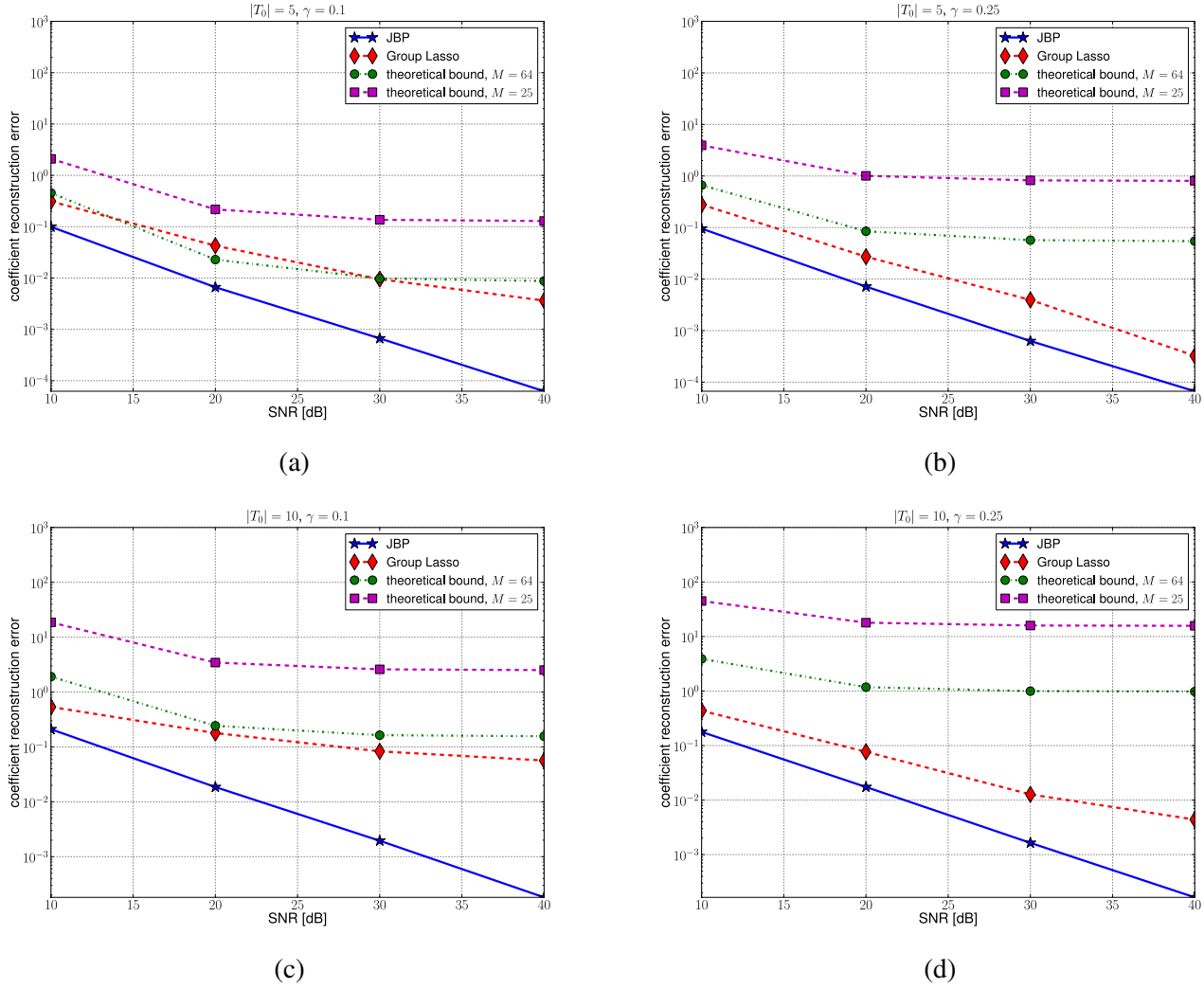


Fig. 2. JBP model recovery performance for random signals. Average relative coefficient reconstruction error is plotted for different signal-to-noise (SNR) ratios between sparse signals and Gaussian noise. Different graphs are for different parameter values: (a) $|T_0| = 5, \gamma = 0.1$, (b) $|T_0| = 5, \gamma = 0.25$, (c) $|T_0| = 10, \gamma = 0.1$, (d) $|T_0| = 10, \gamma = 0.25$.

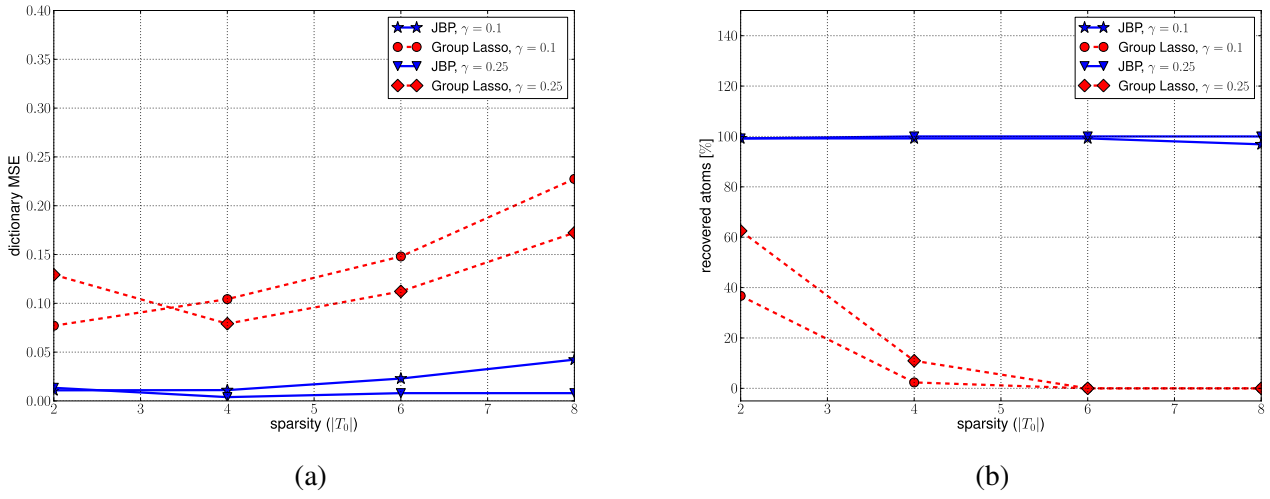


Fig. 3. Recovery performance of dictionary learning using JBP and comparison to GL. (a) Mean square error between recovered atoms and original atoms vs sparsity $|T_0|$, for different values of γ. (b) Percentage of recovered atoms vs sparsity $|T_0|$.

B. Intensity-depth dictionary learning

In our second set of experiments we have evaluated the performance of JBP and dictionary learning on real data, in particular on depth-intensity images. We have learned a depth-intensity overcomplete dictionary on the Middlebury 2006 benchmark depth-intensity data [20]. The intensity data has been whitened, i.e., its frequency spectrum has been flattened, as initially proposed in [5]. Such pre-processing speeds up the learning. Depth data could not be whitened because it would introduce Gibbs artifacts around the missing regions at occlusions. We handle such missing pixels by masking. Learning has been performed in a patch-mode. Namely, in each iteration of the two-step learning process, a large number of depth-intensity pairs of 12×12 patches have been randomly selected from data. Each depth and intensity patch within a pair coincide in a 3D scene. Patches have been normalized to have norm one, and η has been set to 0.1. We have chosen this value such that we get a good reconstruction of depth, without the quantization effects present in Middlebury depth maps (i.e., such that the quantization error is subsumed by the reconstruction error). We have learned dictionaries $\mathbf{\Phi}^I$ and $\mathbf{\Phi}^D$, each of size 144×288, i.e., twice overcomplete. For comparison, we have also learned depth-intensity dictionaries using GL-based learning, where $\lambda=0.3$ has been chosen such that we obtain in average the same reconstruction error as in JBP.

Figures 4(a) and (b) show dictionaries learned by JBP and GL, respectively. The JBP-learned dictionary contains more meaningful features, such as coinciding depth-intensity edges, while GL-learned dictionary only has few of those. JBP dictionary atoms also exhibit correlation between orientations of the Gabor-like intensity atoms and the gradient angle of depth atoms. This is quite visible in the scatter plots of intensity orientation vs depth gradient angle shown in Figure 5. We can see that for JBP there is significant clustering around the diagonal (corresponding to a 90° angle between orientation and gradient), while there is no such effect for GL.

Finally, we have compared the performance of JBP and GL, and the corresponding learned dictionaries, on an inpainting task. Namely, we have randomly removed 95% of depth pixels from an intensity-depth pair obtained by a time-of-flight (TOF) camera[2]. We have chosen the TOF data to show that learned dictionaries of intensity-depth are not linked to particular depth sensors. From the original intensity image and 5% of depth pixels, we have reconstructed the whole depth image, using: 1) JBP and JBP-learned dictionaries $(\mathbf{\Phi}_1^I,\mathbf{\Phi}_1^D)$, 2) GL and GL-learned dictionaries $(\mathbf{\Phi}_2^I,\mathbf{\Phi}_2^D)$, and 3) GL and JBP-learned dictionaries $(\mathbf{\Phi}_1^I,\mathbf{\Phi}_1^D)$. We can see from the Figure 6 that 1) gives the best performance, followed by 3) and 2). Thus, both the JBP dictionary and the JBP inference result in much better performance compared to GL. We do not present comparisons with methods based on depth only (e.g. [8]) since their performance is inferior as they do not use the intensity information.

VII. Conclusion

We have presented an algorithm to learn joint overcomplete dictionaries of image intensity and depth. The proposed method is based on a novel second order cone program for recovering sparse signals of joint sparse support in dictionaries with two modalities, named JBP. We have derived a theoretical bound for the coefficient recovery error of JBP and shown the superiority of JBP to the Group Lasso algorithm through numerical simulations. When applied to the Middlebury image-depth database, the proposed learning algorithm converges to an overcomplete dictionary of interesting intensity-depth features, such as coinciding edges and image grating - depth slant pairs. JBP and this learned dictionary yield state of the art performance in depth inpainting. Both JBP and our learned dictionary are of important value to 3D technologies based on hybrid image-depth sensors.

Appendix

A. Proof of Theorem 1

Before proving this theorem, let us prove the following lemma.

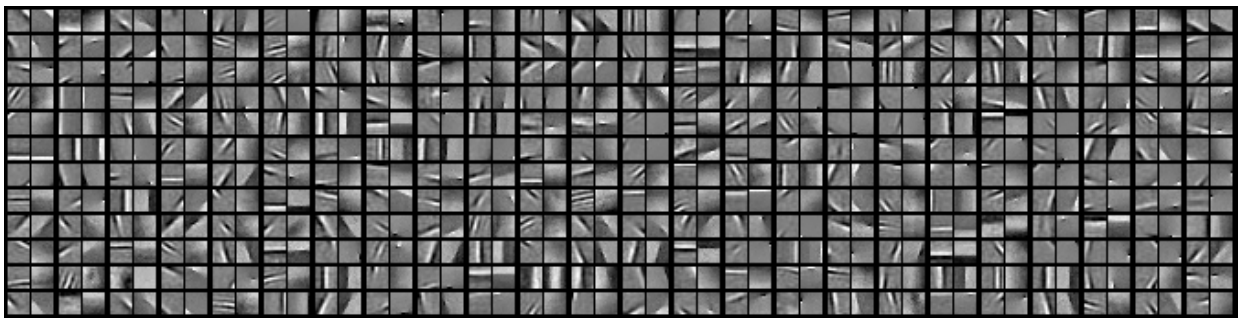
Lemma 3. *For*

$$\mathbf{h}:=[\mathbf{a}^*;\mathbf{b}^*]-[\mathbf{a}^0;\mathbf{b}^0]$$

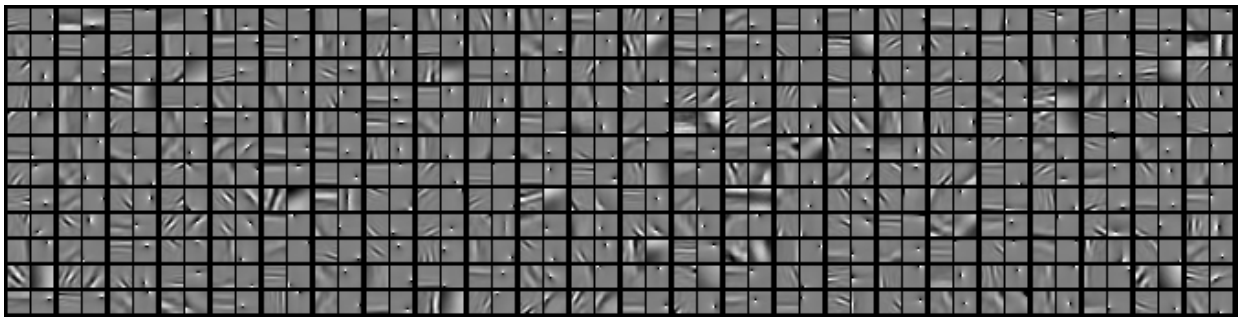
it holds true that:

$$\|\mathbf{h}_{T_0^C}\|_1\leq\|\mathbf{h}_{T_0}\|_1+\gamma U|T_0|, \tag{32}$$

[2]http://www.pmdtec.com/

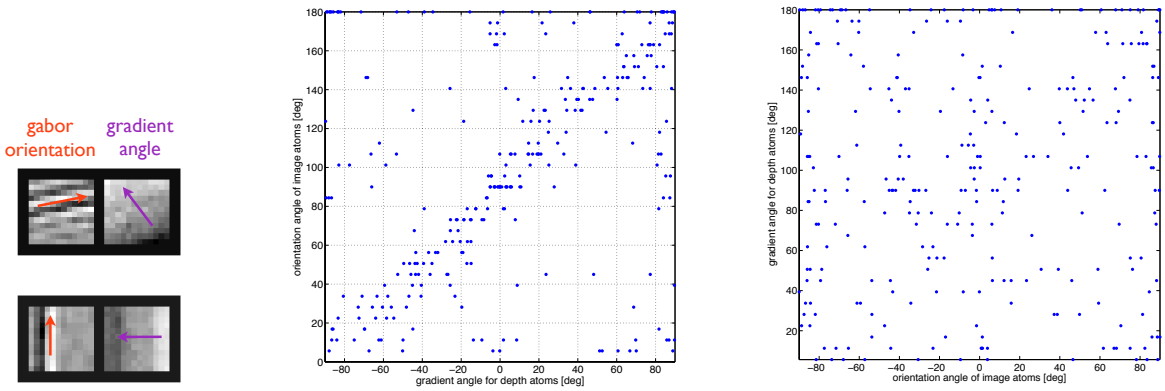


(a) JBP



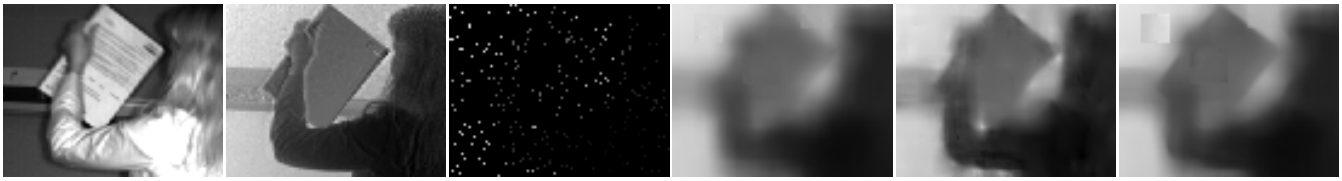
(b) GL

Fig. 4. Learned intensity-depth dictionaries. Each column contains a set of atom pairs (ϕ_1^I, ϕ^D), where the left part is an intensity atom and the right part is a depth atom. (a) JBP-learned dictionaries $(\mathbf{\Phi}_1^I, \mathbf{\Phi}_1^D)$, (b) GL-learned dictionaries $(\mathbf{\Phi}_2^I, \mathbf{\Phi}_2^D)$.



(a) (b) (c)

Fig. 5. Correlation between depth atom gradients and image atom orientations. a) Illustration of atom pairs that have 90 degrees angle between the orientation of the Gabor-like intensity part and the gradient angle of the depth part. Scatter plots of orientation vs gradient angle for b) JBP and c) GL.



(a) (b) (c) (d) GL 25.7 dB (e) JBP 27.8dB (f) GL-$\mathbf{\Psi}_1$ 26.6dB

Fig. 6. Inpainting results on time of flight data. a) Original intensity image, b) Original depth image, c) 5% of kept depth pixels, d) reconstructed depth with GL and GL-learned dictionaries $(\mathbf{\Phi}_2^I, \mathbf{\Phi}_2^D)$, e) reconstructed depth with JBP and JBP-learned dictionaries $(\mathbf{\Phi}_1^I, \mathbf{\Phi}_1^D)$, f) reconstructed depth with GL and JBP-learned dictionaries $(\mathbf{\Phi}_2^I, \mathbf{\Phi}_2^D)$.

where T_0^C denotes the complement set of T_0 and $\mathbf{h}_T$ denotes the subvector of $\mathbf{h}$ corresponding to T.

Proof: Define

$$\begin{aligned}\mathcal{I}_a^0 :=& \{i \in \mathcal{I} : |a_i^0| = U{x_i}^0\},\\ \mathcal{I}_b^0 :=& \{i \in \mathcal{I} \setminus \mathcal{I}_a^0 : |b_i^0| = U{x_i}^0\},\\ \mathcal{I}_a^* :=& \{i \in \mathcal{I} : |a_i^*| = U{x_i}^*\},\\ \mathcal{I}_b^* :=& \{i \in \mathcal{I} \setminus \mathcal{I}_a^* : |b_i^*| = U{x_i}^*\}.\end{aligned}$$

Due to Lemma 2, we have that $\mathcal{I}_a^0 \cup \mathcal{I}_b^0 = \mathcal{I}$ and $\mathcal{I}_a^* \cup \mathcal{I}_b^* = \mathcal{I}$, and due to the definition above it holds that $\mathcal{I}_a^0 \cap \mathcal{I}_b^0 = \emptyset$ and $\mathcal{I}_a^* \cap \mathcal{I}_b^* = \emptyset$. Therefore, we have that:

$$\begin{aligned}\|[\mathbf{a}^*;\mathbf{b}^*]\|_1 &= \sum_{i\in\mathcal{I}_a^*}|a_i^*| + \sum_{i\in\mathcal{I}_b^*}|b_i^*| + \sum_{i\in\mathcal{I}_a^*}|b_i^*| + \sum_{i\in\mathcal{I}_b^*}|a_i^*|\\ &\le U\sum_{i\in\mathcal{I}}|x_i^*| + U\sum_{i\in\mathcal{I}_a^*}|x_i^*| + U\sum_{i\in\mathcal{I}_b^*}|x_i^*|\\ &= 2U\|\mathbf{x}^*\|_1. \end{aligned} \tag{33}$$

Similarly, we have that:

$$\begin{aligned}\|[\mathbf{a}^0;\mathbf{b}^0]\|_1 &= \sum_{i\in\mathcal{I}_a^0}|a_i^0| + \sum_{i\in\mathcal{I}_b^0}|b_i^0| + \sum_{i\in\mathcal{I}_a^0}|b_i^0| + \sum_{i\in\mathcal{I}_b^0}|a_i^0|\\ &\ge U\sum_{i\in\mathcal{I}}|x_i^0| + \min_{i\in T_0}\alpha_i\Big(\sum_{i\in\mathcal{I}_a^0}|a_i^0| + \sum_{i\in\mathcal{I}_b^0}|b_i^0|\Big)\\ &\ge^{(18)} 2U\|\mathbf{x}^0\|_1 - \gamma U|T_0|. \end{aligned} \tag{34}$$

Due to optimality of $\mathbf{x}^*$, we have $\|\mathbf{x}^*\|_1 \le \|\mathbf{x}^0\|_1$, which combined with (33) and (34) gives:

$$\|[\mathbf{a}^*;\mathbf{b}^*]\|_1 \le 2U\|\mathbf{x}^0\|_1 \le \|[\mathbf{a}^0;\mathbf{b}^0]\|_1 + \gamma U|T_0|. \tag{35}$$

Due to $\mathbf{a}_{T_0^C}^0 = \mathbf{0}$ and $\mathbf{b}_{T_0^C}^0 = \mathbf{0}$, we can write

$$\begin{aligned}\|[\mathbf{a}^0;\mathbf{b}^0] + \mathbf{h}\|_1 &= \|[\mathbf{a}_{T_0}^0;\mathbf{b}_{T_0}^0;\mathbf{0}] + [\mathbf{h}_{T_0};\mathbf{h}_{T_0^C}]\|_1\\ &= \|[\mathbf{a}_{T_0}^0;\mathbf{b}_{T_0}^0] + \mathbf{h}_{T_0}\|_1 + \|\mathbf{h}_{T_0^C}\|_1. \end{aligned} \tag{36}$$

Thus, using the triangle inequality and the definition of $\mathbf{h}$ we derive:

$$\|[\mathbf{a}^0;\mathbf{b}^0]\|_1 - \|\mathbf{h}_{T_0}\|_1 + \|\mathbf{h}_{T_0^C}\|_1 \le \|[\mathbf{a}^0;\mathbf{b}^0] + \mathbf{h}\|_1 = \|[\mathbf{a}^*;\mathbf{b}^*]\|_1 \le^{(35)} \|[\mathbf{a}^0;\mathbf{b}^0]\|_1 + \gamma U|T_0|$$

and thus

$$\|\mathbf{h}_{T_0^C}\|_1 \le \|\mathbf{h}_{T_0}\|_1 + \gamma U|T_0|. \tag{37}$$

■

We are now ready to prove Theorem 1.

Proof: Let $\mathbf{A}$ be defined as in Eq. (12). Then we have from (7) and (8) that

$$\|\mathbf{A}\mathbf{h}\|_2 \le 4\epsilon = 4\eta f_0.$$

Assume we have divided T_0^C into subsets of size M, more precisely, we have $T_0^C = T_1 \cup \cdots \cup T_{n-|T_0|}$, where T_i are sorted by decreasing order of $\mathbf{h}_{T_0^C}$, and where $T_{01} = T_0 \cup T_1$. Without alternations - cf. [11] - it holds true that

$$\|\mathbf{h}_{T_{01}^C}\|_2^2 \le \|\mathbf{h}_{T_0^C}\|_1/M.$$

Using (37) yields now

$$\begin{aligned}\|\mathbf{h}_{T_{01}^C}\|_2^2 &\le (\|\mathbf{h}_{T_0}\|_1 + \gamma U|T_0|)^2/M\\ &\le (|T_0|\|\mathbf{h}_{T_0}\|_2 + \gamma U|T_0|)^2/M, \end{aligned} \tag{38}$$

where the second step follows from the norm inequality. Hence:

$$\begin{aligned}\|\mathbf{h}\|_2^2 &= \|\mathbf{h}_{T_{01}}\|_2^2 + \|\mathbf{h}_{T_{01}^C}\|_2^2 \\ &\leq (1+\frac{|T_0|^2}{M})\|\mathbf{h}_{T_0}\|_2^2 + \frac{2\gamma U|T_0|^2}{M}\|\mathbf{h}_{T_0}\|_2 + \frac{(\gamma U|T_0|)^2}{M}. \end{aligned} \tag{39}$$

From the restricted isometry hypothesis, cf. Def. 1, we get

$$\begin{aligned}\|\mathbf{A}\mathbf{h}\|_2 &= \|\mathbf{A}_{T_{01}}\mathbf{h}_{T_{01}} + \sum_{j\geq 2}\mathbf{A}_{T_j}\mathbf{h}_{T_j}\|_2 \\ &\geq \|\mathbf{A}_{T_{01}}\mathbf{h}_{T_{01}}\|_2 - \|\sum_{j\geq 2}\mathbf{A}_{T_j}\mathbf{h}_{T_j}\|_2 \\ &\geq \|\mathbf{A}_{T_{01}}\mathbf{h}_{T_{01}}\|_2 - \sum_{j\geq 2}\|\mathbf{A}_{T_j}\mathbf{h}_{T_j}\|_2 \\ &\geq \sqrt{1-\delta_{M+|T_0|}}\|\mathbf{h}_{T_{01}}\|_2 - \sqrt{1+\delta_M}\sum_{j\geq 2}\|\mathbf{h}_{T_j}\|_2 \\ &\geq \sqrt{1-\delta_{M+|T_0|}}\|\mathbf{h}_{T_0}\|_2 - \sqrt{1+\delta_M}\sum_{j\geq 2}\|\mathbf{h}_{T_j}\|_2 \end{aligned} \tag{40}$$

where δ_S is a constant chosen such that the inequalities hold, which follows from inequality (4) in [11]. Here, $\mathbf{A}_T$ denotes the columns of $\mathbf{A}$ corresponding to the index set T.

In analogy to [11], due to the ordering of the sets T_j by decreasing order of coefficients, we have:

$$|\mathbf{h}_{T_{j+1}(t)}| \leq \|\mathbf{h}_{T_j}\|_1/M$$

meaning each component in $\mathbf{h}_{T_{j+1}}$ is smaller than the average of the components in $\mathbf{h}_{T_j}$ (absolute value-wise). Thus, we get:

$$\begin{aligned}\|\mathbf{h}_{T_{j+1}}\|_2^2 &= \sum_{t\in T_{j+1}}\|\mathbf{h}_t\|_2^2 \\ &\leq \sum_{t\in T_{j+1}}\|\mathbf{h}_{T_j}\|_1^2/M^2 \\ &\leq M\|\mathbf{h}_{T_j}\|_1^2/M^2 = \|\mathbf{h}_{T_j}\|_1^2/M, \end{aligned}$$

and

$$\begin{aligned}\sum_{j\geq 2}\|\mathbf{h}_{T_j}\|_2 &\leq \sum_{j\geq 1}\|\mathbf{h}_{T_j}\|_1/\sqrt{M} \\ &= \|\mathbf{h}_{T_0^C}\|_1/\sqrt{M} \\ &\leq^{(37)} (\|\mathbf{h}_{T_0}\|_1 + \gamma U|T_0|)/\sqrt{M} \\ &\leq \sqrt{|T_0|/M}\|\mathbf{h}_{T_0}\|_2 + \gamma U|T_0|/\sqrt{M} \end{aligned} \tag{41}$$

where the last step follows from the norm inequality. Combining Eq. (41) and Eq. (40), we get:

$$\|A\mathbf{h}\|_2 \geq \sqrt{1-\delta_{M+|T_0|}}\|\mathbf{h}_{T_0}\|_2 - \sqrt{1+\delta_M}\sqrt{|T_0|/M}\|\mathbf{h}_{T_0}\|_2 - \gamma U|T_0|\sqrt{1+\delta_M}/\sqrt{M} \tag{42}$$

and subsequently:

$$\|\mathbf{h}_{T_0}\|_2 \leq \frac{\|A\mathbf{h}\|_2 + \gamma U|T_0|\sqrt{1+\delta_M}/\sqrt{M}}{\sqrt{1-\delta_{M+|T_0|}} - \sqrt{1+\delta_M}\sqrt{|T_0|/M}} \tag{43}$$

$$\begin{aligned}&\leq \frac{4\eta f_0\sqrt{M} + \gamma f_0|T_0|\sqrt{1+\delta_M}}{\sqrt{M(1-\delta_{M+|T_0|})} - \sqrt{|T_0|(1+\delta_M)}} \\ &= Cf_0, \end{aligned} \tag{44}$$

if the denominator is greater than zero. Replacing this result in Eq. (39) and taking $U = f_0$ we get:

$$\|\mathbf{h}\|_2^2 \leq (1 + \frac{|T_0|^2}{M})C^2 f_0^2 + 2\gamma \frac{|T_0|^2}{M} C f_0^2 + \gamma^2 \frac{|T_0|^2}{M} f_0^2, \tag{45}$$

which is equivalent to (19) and thus completes the proof. ■

REFERENCES

[1] D. Hagebeuker, "A 3d time of flight camera for object detection," in *Optical 3-D Measurement Techniques*, 2007.

[2] T. Oggier, M. Lehmann, R. Kaufmann, M. Schweizer, M. Richter, P. Metzler, G. Lang, F. Lustenberger and N. Blanc, "An all-solid-state optical range camera for 3D real-time imaging with sub-centimeter depth resolution (SwissRanger)," in *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, vol. 5249, 2004, pp. 534–545.

[3] Microsoft, "http://www.xbox.com/kinect."

[4] A. Kubota, A. Smolic, M. Magnor, M. Tanimoto, T. Chen and C. Zhang, "Multiview Imaging and 3DTV," *IEEE Signal Processing Magazine*, vol. 24, no. 6, pp. 10–21, 2007.

[5] B.A. Olshausen and D. Field, "Sparse coding with an overcomplete basis set: A strategy employed by V1?" *Vision Research*, vol. 37, no. 23, pp. 3311–25, 1997.

[6] M.S. Lewicki and T.J. Sejnowski, "Learning overcomplete representations," *Neural Computation*, vol. 12, no. 2, pp. 337–65, 2000.

[7] M. Elad and M. Aharon, "Image denoising via sparse and redundant representations over learned dictionaries," *IEEE Transactions on Image Processing*, vol. 15, no. 12, pp. 3736 –3745, 2006.

[8] I. Tošić, B.A. Olshausen and B.J. Culpepper, "Learning sparse representations of depth," *IEEE Journal on Selected Topics in Signal Processing*, vol. 5, no. 5, pp. 941 – 952, 2011.

[9] S. Bakin, "Adaptive Regression and Model Selection in Data Mining Problems," *unpublished doctoral thesis, Australian National University*, 1999.

[10] E. J. Candés and T. Tao, "Decoding by linear programming," *IEEE Transactions on Information Theory*, vol. 51, no. 12, pp. 4203 – 4215, 2005.

[11] E. J. Candès, J. Romberg, and T. Tao., "Stable signal recovery from incomplete and inaccurate measurements," *Communications on Pure and Applied Mathematics*, vol. 59, no. 8, pp. 1207–1223, 2006.

[12] D. L. Donoho, "For Most Large Underdetermined Systems of Equations, the Minimal ℓ^1-norm Near-Solution Approximates the Sparsest Near-Solution," *Communications on Pure and Applied Mathematics*, vol. 59, no. 7, pp. 907–934, 2006.

[13] S. Jokar and M. E. Pfetsch, "Exact and Approximate Sparse Solutions of Underdetermined Linear Equations," *SIAM Journal on Scientific Computing*, vol. 31, no. 1, pp. 23–44, 2008.

[14] T. Tsuchiya, "A convergence analysis of the scaling-invariant primal-dual path-following algorithms for second-order cone programming," *Optimization Methods and Software*, vol. 11, pp. 141–182, 1998.

[15] E. D. Andersen, C. Roos, and T. Terlaky, "On implementing a primal-dual interior-point method for conic quadratic optimization," *Mathematical Programming*, vol. 95, no. 2, pp. 249–277, 2003.

[16] CPLEX, "IBM ILOG CPLEX Optimizer," http://www-01.ibm.com/software/integration/optimization/cplex-optimizer/.

[17] B. A. Turlach, W. N. Venables, and S. J. Wright, "Simultaneous Variable Selection," *Technometrics*, vol. 47, no. 3, pp. 349–363, 2005.

[18] J. A. Tropp, "Algorithms for simultaneous sparse approximation. Part II: Convex relaxation," *Signal Processing*, vol. 86, no. 3, pp. 589 – 602, 2006.

[19] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 68, no. 1, pp. 49–67, 2006.

[20] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *International Journal of Computer Vision*, vol. 47, no. (1/2/3), pp. 7–42, 2002.